Enhanced Semantic Search with Multi-Tag Analysis using BERT

Aditya Rai
Computer Science
Vellore Institute of Technology
Chennai
Tamil Nadu,India
aditya.rai2021@vitstudent.ac.in

Mohamed Naveed
Computer Science
Vellore Institute of Technology
Chennai
Tamil Nadu,India
mohamed.naveed2021@vitstudent.ac.in

Sujithra Kanmani R
Computer Science
Vellore Institute of Technology
Chennai
Tamil Nadu,India
sujithrakanmani.r@vit.ac.in

Abstract— In inner infiite loops of text content, at times and somewhat abnormally happening, semantic search is proving its necessity in infomation retrival systems, considering consistent and even more appropriate search results. The paper investigates the use of Bidirectional Encoder Representations from Transformers (BERT) models for tag generation in semantinc search. The experiment involves building a Named Entity Recognition (NER) model from scratch and training it on a dataset with the aim of classifying search text into respective entities. After this process, BERT based models are called upon to create tags using the entities that have been located. Those tags are then used to search for appropriate places or documents that contain the relevant information. The methodology includes multi-tags analysis where the common data figures applications of the tags are interrelated to improve relevance of the results. The research depicts how BERT based semantic search contributed to better information retrieval in terms of accuracy and relevance. The experimental results showed that the system can work with difficult queries and gives more relevant search results.

Keywords— Semantic Search, BERT, Tag Generation, Keyword Extracion, Contextual Understanding, Deep Learning, Information Retrieval

Introduction

Efficient information retrieval is vital in the digital age. Traditional keyword-based search often falls short in capturing the intent and context behind user queries, leading to the rise of semantic search, which focuses on understanding query intent and retrieving contextually relevant results.

Semantic search utilizes natural language processing (NLP) and machine learning (ML), with models like Bidirectional Encoder Representations from Transformers (BERT) playing a key role. Introduced by Google in 2018, BERT captures bidirectional contextual information from large text datasets. This paper explores using BERT in semantic search to improve tag generation—a critical component for organizing information and enhancing search relevance. By leveraging BERT's semantic understanding, we aim to improve the accuracy of search outcomes.

Our methodology involves training a Named Entity Recognition (NER) model to identify entities in queries, forming a foundation for BERT-based tag generation. This enables tags that capture not only keywords but also semantic context.

Additionally, multi-tag analysis helps identify intersections between tags, enhancing search precision. For example, if a query includes entities related to technology and healthcare, the system retrieves documents linking both fields for a more comprehensive result.

This research addresses challenges like ambiguity and polysemy in natural language queries, with experimental results validating the effectiveness of BERT-based tag generation for improved search relevance.

I. Literature Survey

A. Advancements in Information Extraction and NLP

DBXplorer[1] marks a forward leap in database systems especially in keyword based searching in relational databases. This novel model, DBXplorer, is different from traditional database querying techniques in that it has a simple interface that makes the access to data practices straightforward and easy without having to comprehend the database structure in detail. The scope of DBXplorer however does not only stop at database management and database-driven applications but may in fact extend to video content management systems by offering means of optimizing content organization and access of large data networks.
Also, with the introduction of dual decoder[3] models for natural language processing, the searching has gradually elevated to another level. These types of models are great at the translation of natural language queries to keyword queries solving the problem of lexical gap for better information retrieval. Such Improvements are very useful in tagging of contents and in searching for them which guarantees a better and more muscular searching system.
The use of BERT variants[6] for keyword extraction is another sophisticated method for sectioning relevant data from any form of text. They apply advanced techniques of natural language processing to spot critical sentences that aid in summarization and arrangement of content. The BERT architecture based models' content management capabilities are very promising – the possibilities of changing content and metadata creation and usage are there, thus enhancing content searching in the quality and expeditiousness that is unprecedented'[4][7].

Considerable advances in the accuracy and interpretability of Named Entity Recognition (NER) systems have been achieved through the implementation of hybrid NER approaches that rely on machine learning, deep learning and/or the hand-coding of rules [8]. It is not sufficient for tagging and indexing, as this factor includes understanding what the content carries more than it classifies. Thus, facilitating a cognitive-based approach in analyzing and retrieving of video materials rather than mere classificatory retrieval of such materials. [9]

B. *Keyword Extraction and SEO in Content Discoverability*

For instance, the ease and effectiveness of methods such as RAKE – which stands for Rapid Automatic Keyword Extraction – in keyword identification all underscore the importance of metadata creation in promoting the visibility of content [2]. These methodologies, in particular, offer an insight of such a high level of textual understanding that it can turn out to be very beneficial in searching for video content, thus, increasing the usability of such content to the users.[5]

Additionally, the recent developments in keyword generation strategies in search engine advertising assert that, apart from the need for economically viable keywords, there is the need for keywords that are relevant. These allow a refined form of digital marketing in that there is a compromise between the costs incurred and the quality of traffic directed to the website. These strategies in turn in video content marketing can increase the reach and the engagement at very minimal cost for the producers and the advertisers as well. [10][13]

Also, the evaluation of the techniques known as Search Engine Optimization (SEO) and the Influence of these techniques on the visibility of web content, helps in explaining how keywords can be strategically used [11]. Where video content is concerned, the use of strategized keywords to improve organic reach and engagement serves the purpose of improving the online presence of the video content. All of this demonstrates just how important SEO is when it comes to devising any of the digital content strategies and brings out the necessity of this subject to this study that revolves around tag generation for video content management using BERT for semantic search [12].

C. *Semantic Search Systems and Future Directions*

A semantic search system research contributes to understanding the dynamics of information retrieval today and tomorrow. These include focus on entities, relationships, and contextual knowledge in order to provide a search experience that is more natural and human like. The use of semantic search solutions in video content management systems would improve the retrieval of content making it enticing to users – a revolutionary way of interacting and consuming digital contents[14].

II. *Materials and Methods*

A. *Dataset Description*

The CoNLL-2003 dataset is a standard dataset in NLP for named entity recognition (NER) research. It was collected from the Reuters Corpus and first appeared at CoNLL in 2003. It has also detailed named entity tags such as person, organization and location. The data is provided in a very clearly set out table format allowing for in-depth analysis and training of models. It is used by the researchers to build NER systems that detect the entities from the text automatically. Some of these challenges are class imbalance and udget naming. The three-way division of data into training, validation, and test sets is done to avoid bias in the evaluation of the model. It is very important for the development of NER systems as well as other NLP tasks.

B. *Dataset Analysis*

Exploratory Data Analysis (EDA): The CoNLL dataset consists of natural language text which has been tokenized and linguistically annotated, for instance, providing information about the parts of speech or named entities. There are rules that govern the manner in which such data is presented for easy processing: for instance, the CoNLL format that separates the data in columns using tab delimiters. These annotations indicate the features that the model will be trained to learn while training the model. The volume of the trial data is quite different, beginning with thousands and up to millions of examples. Some quality control mechanisms check for errors and consistency in judgments. There may be human annotation in the process of creating such data. Such data is important in training, for example, machine learning models which perform named entity recognition and part of speech tagging.

The data in a CoNLL dataset is a compilation of annotated text that is used for training machine learning systems to perform various NLP tasks. This is the curriculum for the models that processes text and understands the included linguistic features.

The models learn to obtain the required annotations starting from the text data and modifying the weights to obtain the smallest value of the objective function.

In general, the data is the basis for building models that are capable of comprehending and manipulating linguistic aspects of written language.

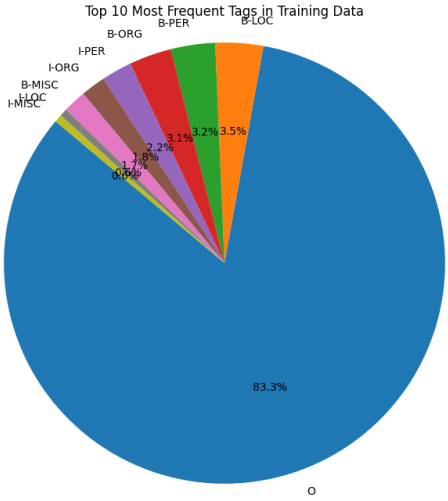

Figure 4: Pie chart representation for all tags in data

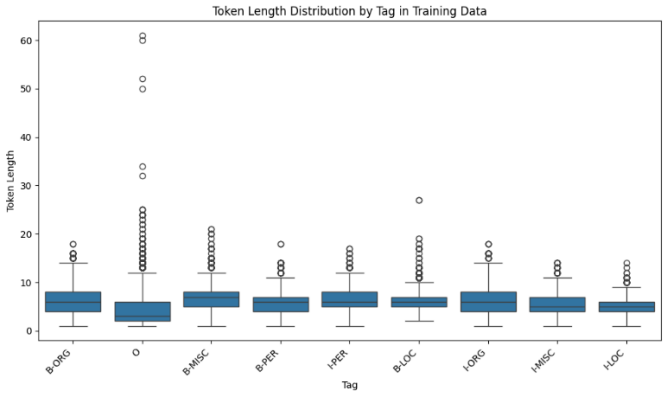

Figure 5: Outliers for token length against tags in training data

C. Model Description

BERT is a state-of-the-art deep neural network model developed by Google; its full form is Bidirectional Encoder Representations from Transformers. One of the common structures of the Transformer's architecture, such models are unprecedented in their efficiency for Natural Language Processing tasks. Unlike earlier applications which processed the text in a certain direction, one-way BERT can read the text from either end and all the while, both ends – thus it is, by nature biconditional.
One of the most crucial aspects of BERT is the approach to pretraining the model; more specifically we mean training on non labeled data at vast scale. This stage being chiefly concerned with the understanding of word level semantics, relations between words in a sentence are very well taken care of in the BERT model.
With BERT, all text is processed in both directions which allows the model to consider the meaning of words based on other words or even sentences in close proximity. This property of BERT becomes very important in tasks such as text classification, question answering, and named entity recognition where the meaning of some phrases in some contexts is very important for obtaining a correct result.
Another feature of BERT is its effectiveness when dealing with input strings of various lengths. This is achieved through the use of attention mechanism and self-attention layers which allows the model to focus on some parts of the input sequence while disregarding the rest which may not be relevant.
BERT has proven effective for a number of popular metrics used for assessing the performance of different NLP models and even more so for those which are oriented on carrying out particular tasks including question answering where BERT exceeds the expectation for the result that a human can deliver.

D. Architecture Diagram

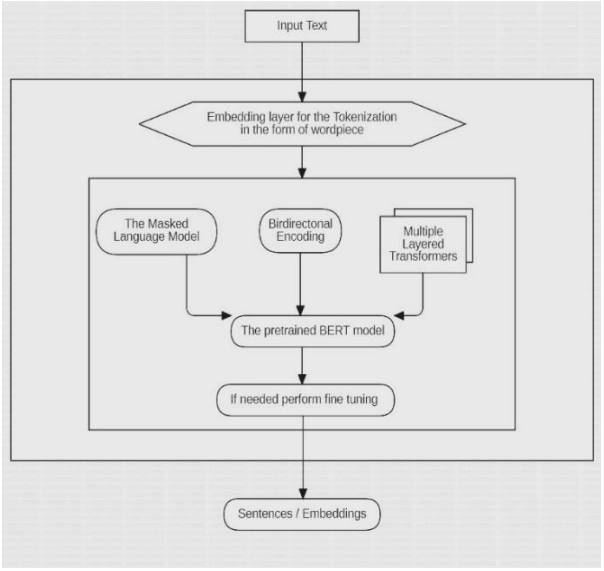

Figure 6: The Architecture of BERT Model

E. Approach Employed

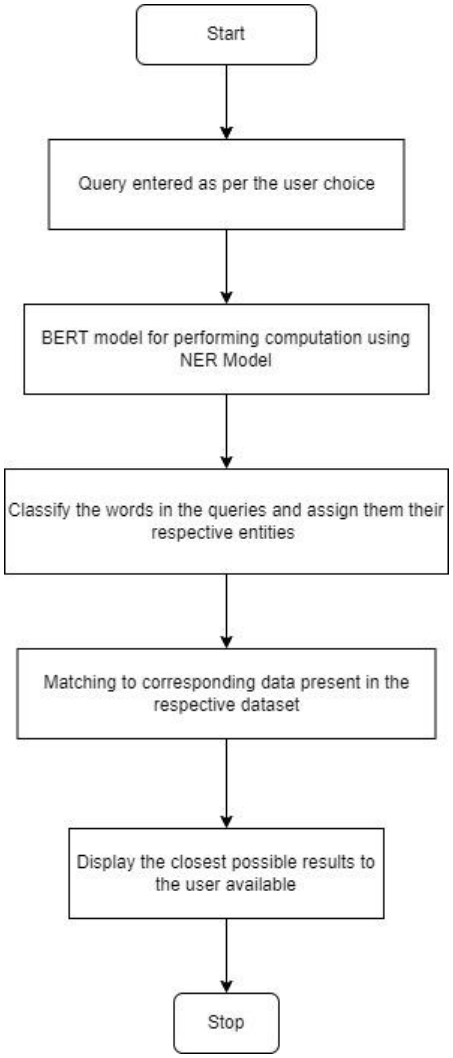

Figure 7: Flowchart of search engine

The technique we adopt for semantic searching and tag creation is based on a complex logic that manages to enjoy the benefits of both BERT-based Named Entity Recognition and understanding the context. The logic applied can be outlined as follows in major areas.
At the center of the system we propose to develop is a flexible and appropriate keyword extraction logic. Based on BERT's understanding of the text in a horizontal way, this logic aims by extracting keywords or entities from a search query and input text. Giving importance to the most pertinent keywords based on its relevance, frequency and context of the words used without distorting the actual idea that the user wants to search for fulfills the aim of that logic.
As an enhancement of the extracted keywords, we apply a sophisticated tag generation algorithm. The algorithm takes the keywords that have been derived and creates tags related to the search term in question. In surmounting the hierarchical boundaries of the keywords, the algorithm logic assists in the generation of the tags in such a, which has a higher return during content retrieval concerning the input query thus improving search overall.
Among the important ones remains the logic emphasizing the importance of contextual relevance. We include contextual relevance logic in order to relate the generated tags to descriptions or documents in our dataset. Such logic limits the degree of relevance of content to be retrieved to a user's search query and ensures that such content is in its appropriate context increasing user search experience and satisfaction levels greatly.
Further, what we implement in this regard is an analysis of common data points. This analysis consists of identifying a set of common words or entities present in various tags produced with respect to search queries.

By analyzing common data points, we refine and prioritize search results, further enhancing the relevance and accuracy of content retrieval

III. Evaluation & Resultss

A. *Model Evaluation Metrics*

In order to evaluate the effectiveness of the model, a number of indices based on equations (1) through (4) are utilized, which include precision, recall, accuracy, and the F1-score.
Precision assesses the correctness or accuracy of the positive predictions, and it can be defined as the ratio of true positive predictions to all positive predictions made

Precision = TP / (TP + FP) (1)

Accuracy is a description of how correct a model's predictions are overall, defined as the ratio of correctly predicted examples (true positive and true negative) to all examples.

Accuracy = (TP + TN)/(TP + TN + FP + FN) (2)

The F1 score incorporates both precision and recall in an attempt at a single measure of performance presented in the statistic. It is determined as a weighted average of precision and recall where both components are given the same importance.

F1 Score = 2 * ((1) * (2)) / ((1) +(2)) (3)

Recall or sensitivity is one of the performance metrics of a model denoting the model's capacity to correctly identify all the true positive instances in a given scenario. In other words, it is the ratio of correctly predicted positive observations to all observations in the actual class.

Recall = TP / (TP + FN) (4)

Here, TP represents true positives and FP Tdenotes false positives.

B. *Search Engine Evaluation Metrics*

Precision
Precision assesses the fraction of relevant tags contained in the total number of tags that the search engine outputs. A high precision score signifies that the search engine is capable of effectively removing tags that are not relevant.

Precision=Number of Relevant Tags Retrieved/Total number of Tags retrieved

Recall
In information retrieval, recall denotes the percentage of relevant tags that have been retrieved in relation to the total number of relevant tags that could be retrieved for a given query. The higher the recall value, the more relevant tags the system is able to get with respect to the total number of relevant tags.

Recall=Number of Relevant Tags Retrieved/Total Number of Relevant Tags

F1 Score
The F1 Score is the balanced measure of precision and recall, providing an overall view of the engine's accuracy and completeness in retrieving relevant tags.

F1=2*((Precision*Recall/Precision+Recall))

C. *Results*

Model:

Epoch	Training Loss	Validation Loss	Precision	Recall	F1	Accuracy
1	0.231400	0.061347	0.918958	0.931088	0.924983	0.983049
2	0.046500	0.057054	0.936130	0.946079	0.941078	0.985782
3	0.025600	0.057549	0.940289	0.947757	0.944008	0.986687
4	0.014800	0.061366	0.941366	0.950106	0.945716	0.986910
5	0.010400	0.062363	0.941672	0.949994	0.945815	0.986989

Figure 16: Evaluation Metrics

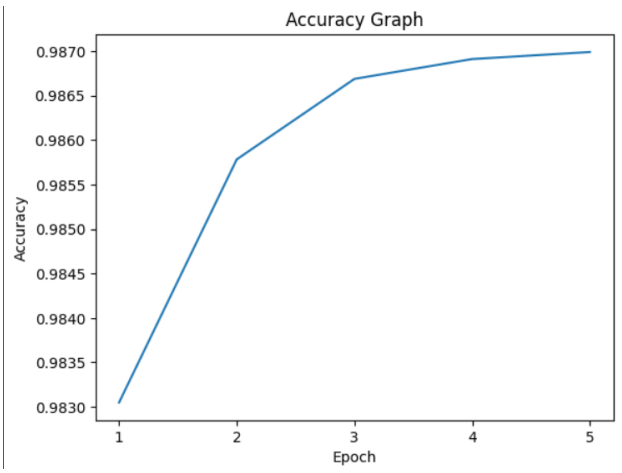

Figure 17: Accuracy Graph

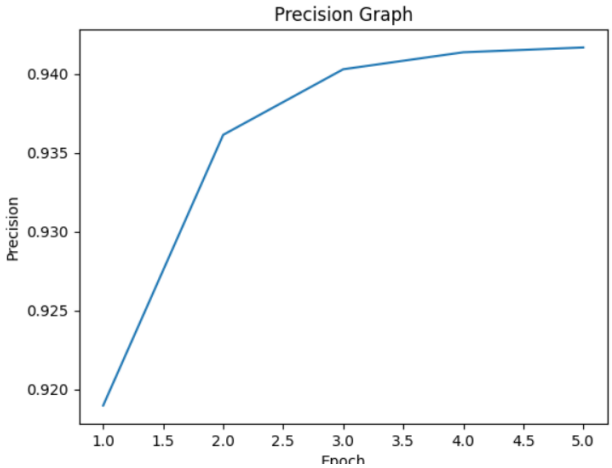

Figure 18: Precision Graph

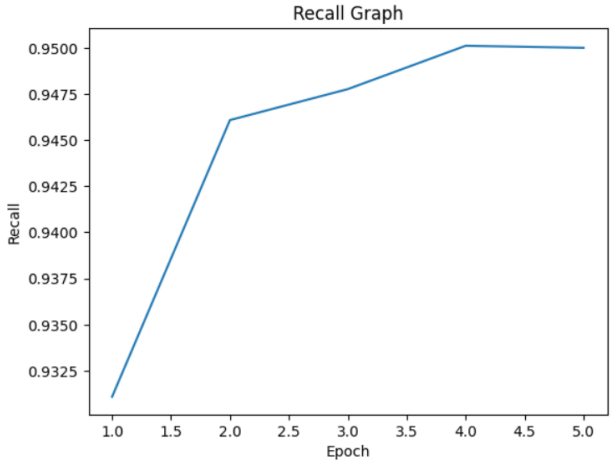

Figure 19: Recall Graph

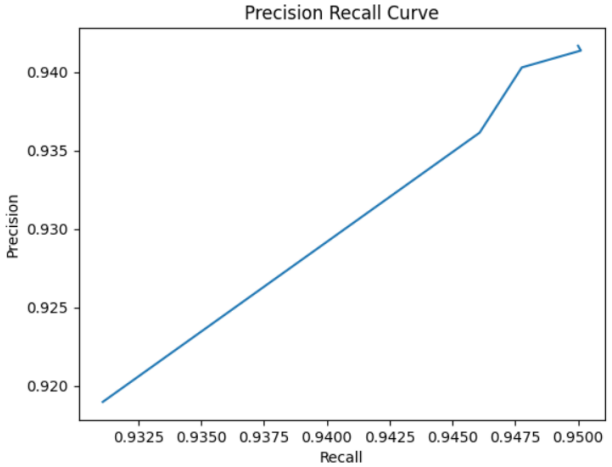

Figure 20: Precision Recall Curve

Search Engine:

```
User 1:
  Average Precision: 0.73
  Average Recall: 0.72
  Average F1 Score: 0.72

User 2:
  Average Precision: 0.85
  Average Recall: 0.81
  Average F1 Score: 0.83

User 3:
  Average Precision: 0.65
  Average Recall: 0.67
  Average F1 Score: 0.66
```

IV. Conclusion

To sum up, our findings showed that BERT-based models can be efficiently used for purposes of semantic search and automatic tagging of video content management systems. With the introduction of new concepts related to Named Entity Recognition, keyword extraction and context, we designed an efficient framework to support better content indexing and information searching.

Our framework which adopts the BERT's bidirectional context understanding, advantages in the tagging of content and more so extracting relevant keywords. The search also applies a logical mapping such that the results are not just random but relevant to what the user is searching for, thus enhancing the search experience.

Moreover, the conducted research embraces the aspects of preprocessing data, training a model, and most importantly retraining the system which is key in performance of semantic searches. The introduction of some performance measures and flexibility of modifying the strategies helps in improving further our semantic search system.

Finally, drawing on the benefits of our study, it is possible to predict development in content management, user experience, and information retrieval. There is still more work to be done where advanced techniques in natural language processing and machine learning will be applied in enhancing the performance of a semantic search system to improve users' experience when looking for or browsing through content.

REFERENCES

[1] Chaudhuri, Surajit, Sanjay Agrawal, and Guatam Das. "System for keyword based searching over relational databases." U.S. Patent 6,801,904, issued October 5, 2004.

[2] Qian, Yili, Chaochao Jia, and Yimei Liu. "BERT-based text keyword extraction." In Journal of Physics: Conference Series, vol. 1992, no. 4, p. 042077. IOP Publishing, 2021.

[3] Mansouri, Alireza, Lilly Suriani Affendey, and Ali Mamat. "Named entity recognition approaches." International Journal of Computer Science and Network Security 8, no. 2 (2008): 339-344.

[4] Nesi, Paolo, Gianni Pantaleo, and Gianmarco Sanesi. "A Distributed Framework for NLP-Based Keyword and Keyphrase Extraction From Web Pages and Documents." In DMS, pp. 155-161. 2015.

[5] Rose, Stuart, Dave Engel, Nick Cramer, and Wendy Cowley. "Automatic keyword extraction from individual documents." Text mining: applications and theory (2010): 1-20.

[6] Beliga, Slobodan, Ana Meštrović, and Sanda Martinčić-Ipšić. "An overview of graph-based keyword extraction methods and approaches." Journal of information and organizational sciences 39, no. 1 (2015): 1-20

[7] Abhishek, Vibhanshu, and Kartik Hosanagar. "Keyword generation for search engine advertising using semantic similarity between terms." In Proceedings of the ninth international conference on Electronic commerce, pp. 89-94. 2007.

[8] Nagpal, Mayank, and J. Andrew Petersen. "Keyword selection strategies in search engine optimization: how relevant is relevance?." Journal of retailing 97, no. 4 (2021): 746-763.

[9] Joshi, Amruta, and Rajeev Motwani. "Keyword generation for search engine advertising." In Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06), pp. 490-496. IEEE, 2006.

[10] Stepanyan, Levon. "Automated custom named entity recognition and disambiguation." International Journal of Signal Processing 5 (2020).

[11] Liu, Xiaoyu, Shunda Pan, Qi Zhang, Yu-Gang Jiang, and Xuanjing Huang. "Generating keyword queries for natural language queries to alleviate lexical chasm problem." In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 1163-1172. 2018.

[12] Ilan, Judit Bar. "Search engine results over time: A case study on search engine stability." Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics 2 (1998): 1.

[13] Ntoulas, Alexandros, Junghoo Cho, and Christopher Olston. "What's new on the Web? The evolution of the Web from a search engine perspective." In Proceedings of the 13th international conference on World Wide Web, pp. 1-12. 2004

[14] Wei, Wang, Payam M. Barnaghi, and Andrzej Bargiela. "Search with meanings: an overview of semantic search systems." International journal of Communications of SIWN 3 (2008): 76-82.